# Designing A Long Lasting Linguistic Project: The Case Study of ASIt

**Maristella Agosti[1], Emanuele Di Buccio[1], Giorgio Maria Di Nunzio[1], Cecilia Poletto[2], Esther Rinke[2]**
1: University of Padua, 2: Goethe Universität Frankfurt Am Main
{agosti,dibuccio,dinunzio}@dei.unipd.it, {poletto,rinke}@em.uni-frankfurt.de

**Abstract**

In this paper, we discuss the requirements that a long lasting linguistic database should have in order to meet the needs of the linguists together with the aim of durability and sharing of data. In particular, we discuss the generalizability of the Syntactic Atlas of Italy (ASIt), a linguistic project that builds on a long standing tradition of collecting and analyzing linguistic corpora, on a more recent project that focuses on the synchronic and diachronic analysis of the syntax of Italian and Portuguese relative clauses.
The results that are presented are in line with the FLaReNet Strategic Agenda that highlighted the most pressing needs for research areas, such as Natural Language Processing, and presented a set of recommendations for the development and progress of Language Resources (LR) in Europe (Soria et al., 2014).

**Keywords:** Linguistic Resources, Digital Libraries, Tagging Relative Clauses

## 1. Background

Language Resources (LRs) are critical in the development of applications for overcoming language barriers, documenting endangered languages, and for supporting research of several fields. The FLaReNet Strategic Agenda highlighted the most pressing needs for research areas, such as Natural Language Processing, and presented a set of recommendations for the development and progress of LRs in Europe (Soria et al., 2014). In particular, the recommendations are organised according to broad thematic clusters that go from availability, sharing and distribution to coverage, quality and adequacy. Language resources that have already been made publicly available vary in the richness of the information they contain (Bird et al., 2009). However, the quality – one of the thematic clusters addressed by the Strategic Agenda – of such corpora may have been reduced by the uncontrolled usage of automatic learning algorithms (Spärck Jones, 2007). Interoperability, sharing and re-use of linguistic resources are issues that are often not carefully addressed in linguistic projects. In fact, the heterogeneity of linguistic projects has been recognized as a key problem limiting the reusability of linguistic tools and data collections (Chiarcos, 2012). A significant example of a study that showed how heterogeneity negatively affects the re-use and integration of data is represented by the Edisyn search engine – the aim of which was to make different dialectal databases comparable – which "in practice has proven to be unfeasible" even when the linguistic framework and the type of data gathered are very similar.[1]
For these reasons, the methodological and technological boundaries existing in each linguistic project need to be overcome in order to find common grounds where linguistic material can be shared and re-used over a long period of time. Consequently, a possibly standardized methodology for designing linguistic databases is necessary to develop linguistic resources that fully meet the desiderata of the Strategic Agenda. Curated databases (Peter Buneman and James Cheney and Wang-Chiew Tan and Stijn Vansummeren, 2008) are a possible solution for designing, controlling and maintaining collections that are consistent, integral and high quality.
One of the first reviews of modern linguistic databases presented a pool of papers that discussed the conceptual modeling of linguistic data, the convenience and efficiency of retrieval, and the availability of technology (König and Mengel, 2000). In addition, linguistic databases have been often used along with highly sophisticated cartographic techniques to increase the knowledge of the spatial distribution of closely related language varieties (Goebl, 2010). This research field called dialectometry has witnessed a great number of improvements from many points of view (Wieling and Nerbonne, 2015). Nevertheless, a detailed conceptual modelling is rarely addressed by the papers presented in these reviews, as these papers focus more on the linguistic aspects rather than on the design of a long lasting curated database.
In this paper, we discuss the requirements that a long lasting linguistic database should have in order to meet the needs of the linguists together with the aim of durability and sharing of data. In particular, we focus on the case study of the Syntactic Atlas of Italy (ASIt) that builds on a long standing tradition of collecting and analyzing linguistic corpora, which has given rise to different efforts and projects on various languages over the years.

## 2. Linguistic Motivations and Requirements

One of the basic problems we have to deal with when setting up a database with linguistic data is related to the qualitatively and quantitatively different types of data that have to be classified and retrieved. A linguistic database with the function of the old linguistic atlases (and hopefully many more) contains in addition to the obvious linguistic data also many other kind of data among those information about geographic locations, the type of inquiries adopted to gather the data, the speakers who have delivered the data, all of them being relevant to the linguistic analysis and therefore to be made accessible to the user. A typical example in which geographic data and linguistic data

[1] http://www.dialectsyntax.org/

interact and have to be both retrieved at the same time is the following: a given linguistic phenomenon is found in the same geographical area as another, or the two linguistic phenomena are in geographically disjunct areas, or the area of the first implies the area of the second. The geographic distribution of phenomena represents precious information for the linguist and should be immediately retrievable from the interface. A second set of data which is important to linguistic research concerns the test subjects used to gather the data on the field. They might provide input to investigate how language changes over time in a given geographical space. Since the data to be combined are of different origin and are to be classified according to different parameters, a careful planning of the structure of the database is necessary to develop a system which has the properties of durability and wide usage among researchers that justify such an expensive enterprise. Each set of data must therefore be analyzed in order to determine which parameters are relevant for its classification; furthermore, each set of data must be related to all others since complex information concerning the interaction of the above mentioned different sets of data must be made accessible.
The need of a careful planning to achieve a long lasting tool has to deal with two factors:

- the fact that each set of data has to be analyzed and a set of facets has to be established according to which the data can be retrieved;
- the fact that all the different sets of data included in the database have to be related to each other.

To provide an example of how difficult it can be to plan a database and show that developing a tag system for the linguistic set of data is by no means a trivial enterprise, let us examine what the annotation of a single set of data requires. Take for instance the purely linguistic data to be annotated: in order to do this a) the research team has to make decisions on linguistic phenomena found in the corpus that are relevant and which ones can be excluded. This means that on the one hand the set of tags has to be enriched with respect to the standard set already used in other databases of the same type. On the other hand, the additional set of tags has to be kept at a minimum, otherwise the user will have a very long and unmanageable list of tags among which to choose. b) The set of tags adopted should be interpretable across linguistic theories and valid across time in order to last as long as possible as old linguistic atlases did. c) When the database is built, it is not always predicable for which purpose it will be used in ten or more years from its release, which means that all the information, even what seems prima facie irrelevant at the moment of planning, should be present and retrievable. The database must have the property of flexibility to be adapted to the various requests which might come up in the future.
A second set of questions to solve has to do with the availability of a search which combines data of different types. Although at the moment our empirical investigation only requires a certain combined search, like for instance the one between linguistic and geographical data mentioned above, in the future it might be needed to cross-check other sets of data (see for instance the above mentioned possibility of using data on the test subjects to investigate different generations of speakers and thus trace the path of linguistic change).
All this points toward the necessity of a careful and rigorous first planning phase which ensures a long lasting and flexible usage.

## 3. The case study of ASIt

In what follows, we present the case of a project whose original design has proved flexible enough to accomodate data coming from different languages and different domains of linguistic research and is currently still being used to store and retrieve data from a new project. The ASIt enterprise builds on a long standing tradition of collecting and analyzing linguistic corpora, which has given rise to different efforts and projects over the years (Benincà and Poletto, 2007; Agosti et al., 2010; Agosti et al., 2011; Agosti et al., 2012). Research on the syntax of Italian is of great interest to several important lines of research in linguistics: it allows comparison between closely related varieties (the dialects), hence the formation of hypotheses about the nature of cross-linguistic parametrization; it allows contact phenomena between Romance and Germanic varieties to be singled out, in those areas where Germanic dialects are spoken; it allows syntactic phenomena of Romance and Germanic dialects to be found, described and analyzed to a great level of detail (Agosti et al., 2012).
The ASIt Digital Library System (DLS) was originally intended to support the first line of research, i.e. comparison between closely related Italian varieties (Agosti et al., 2010). Dialectal data to support this investigation were gathered during a twenty-year-long survey investigating the distribution of several grammatical phenomena across the dialects of Italy (Benincà and Poletto, 2007). The dialectal data is constituted of a set of Italian *questionnaires*, where each questionnaire is constituted of a set of Italian *sentence*s; one or more dialectal questionnaires are associated to an Italian questionnaire, where a dialectal questionnaire is constituted of dialectal translations of the sentences in the Italian questionnaire. Sentence translations were gathered from native speakers from about 200 different locations in Italy. The dialectal data includes information on the speakers, the locations (and administrative divisions at diverse levels), and the morphological and syntactic phenomena occurring in the sentences as well as some non systematic information about the phonology of the dialects. A set of tags was defined to capture syntactic phenomena at sentence level; tags are organized in tag groups, thus leading to a two-level tag hierarchy. The conceptual model of ASIt is depicted in Figure 1. Questionnaires are modeled through the entity `DOCUMENT`.
The corpus to be automatically handled was firstly envisioned and secondly mapped on a conceptual schema in order to be general enough to handle diversified geolinguistic projects with tagging on different linguistic units. This was a crucial methodological investment to support the other lines of research, specifically those involving the relationship between Romance and Germanic varieties and investigated in a multidisciplinary and collaborative project,

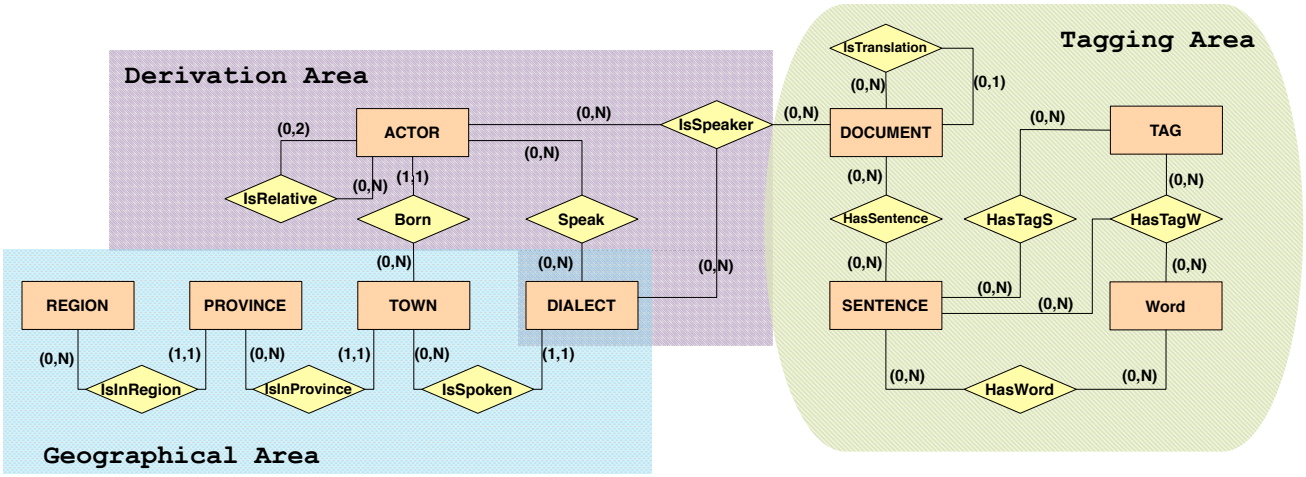


Figure 1: Entity-Relationship Diagram for the ASIt Digital Library.

"Cimbrian as a test case for synchronic and diachronic language variation". In fact, the project aimed at collecting, digitizing and tagging linguistic data from the German variety of Cimbrian, which is spoken in three areas of northern Italy – Giazza (VR), Luserna (TN), and Roana (VI). The dialectal data were constituted of Cimbrian documents spanning over three centuries, each constituted of a set of sentences; documents were treated as ASIt questionnaires. In this project, the phenomena under investigation were also at the word level, differently from ASIt where the unit of interest was only the sentence. A new tag-set was defined to capture word-level phenomena and information on the word constituting the sentences was extracted. Similarly to ASIt, tags were organized in groups according to their linguistic import, but the resulting hierarchy is deeper than the ASIt one.

In order to address the issue of linguistic data and tools reusability, the dialectal data obtained from the ASIt and the Cimbrian linguistic project were exposed through the Linked Open Data paradigm (Di Buccio et al., 2014).

## 4. Discussion and Final Considerations

The generalizability of the ASIt approach that is materialized in the developed conceptual schema has been shown in a recent test case: the DFG-Projekt PO 1642/1-1.[2] The objective of this project is the synchronic and diachronic analysis of the syntax of Italian and Portuguese relative clauses. Since the project aimed at investigating a set of phenomena related to different types of relative clauses, syntactic phenomena under investigation are captured through a new dedicated sentence level tag set tailored for this project. This database is the first attempt to investigate different types of relative clauses in a corpus of spoken colloquial language in a systematic way. The challenge consisted in adapting the tools of the ASIt project to the corpus data, i.e. adapting a design originally crated to deal with purely experimental setting to a much freer and less controlled set of data coming from a pre-existing corpus. Originally, the Asit tools were developed for Italian dialectal data which were gathered by asking speakers to translate standard Italian sentences into their dialect. In the present project, the data consisted of spontaneous and colloquial speech. In addition, the tags used to classify the Italian relative clauses in the ASIt had to be adapted to Portuguese, where other phenomena could potentially be of interest.

Figure 2 reports a screenshot of the web application designed to access the dialectal data by tags (co)occurrence. Tag categories (first level of the tag hierarchy) are reported on the left. When the user clicks on a category (e.g. "Relatives"), the list of tags in that category are displayed; the user can select one or more tags to specify its information need, and specify if tag co-occurrence should be optional (OR) or not (AND). A screenshot of the search result page is reported in Figure 3.

The fact that the original setup has been successfully applied on different languages (German and Portuguese dialects) spoken in different geographical areas and even for data with historical significance shows that its flexibility and coherence were sufficient to serve as a form of reference model to be adopted when transposed to a new set of data.

[2] http://ims.dei.unipd.it/websites/portuguese-relclauses/index.html

## 5. Acknowledgements

This work has been partially supported by the project 'Synchronische und diachronische Analyse der Syntax italienischer und portugiesischer Relativsätze' (DFG-Projekt PO 1642/1-1).

## 6. Bibliographical References

Agosti, M., Benincà, P., Nunzio, G. M. D., Miotto, R., and Pescarini, D. (2010). A Digital Library Effort to Support the Building of Grammatical Resources for Italian Dialects. In Maristella Agosti, et al., editors, *Digital Libraries - 6th Italian Research Conference, IRCDL 2010. Revised Selected Papers*, volume 91 of *Communications in Computer and Information Science*, pages 89–100. Springer.

Agosti, M., Alber, B., Nunzio, G. M. D., Dussin, M., Pescarini, D., Rabanus, S., and Tomaselli, A. (2011).

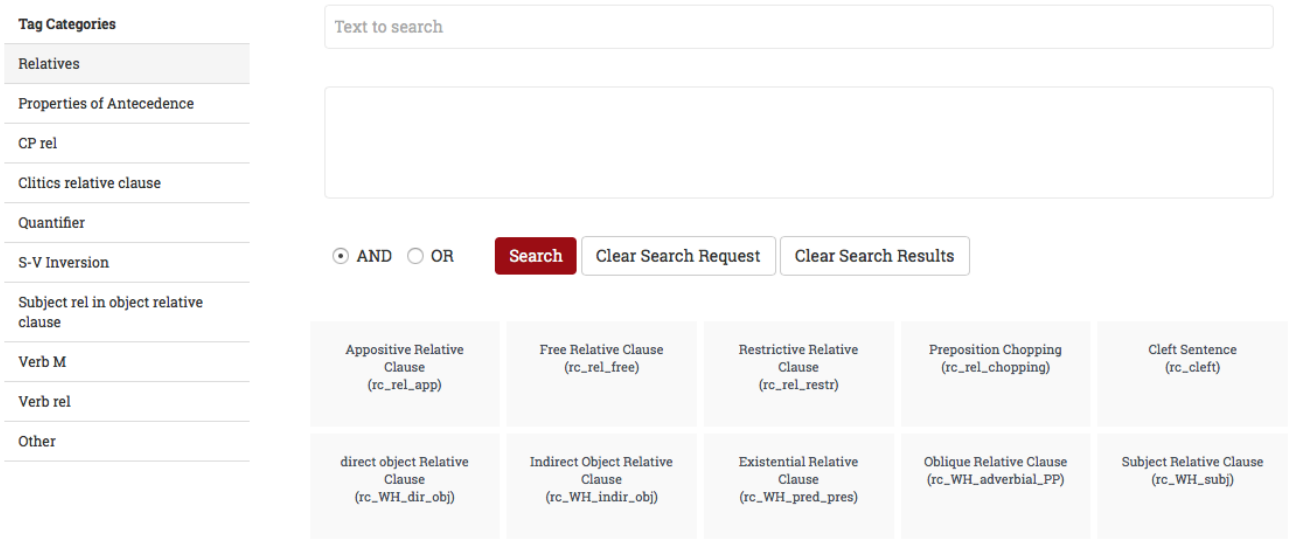


Figure 2: A screenshot of the Web Application designed for the DFG-Projekt PO 1642/1-1.

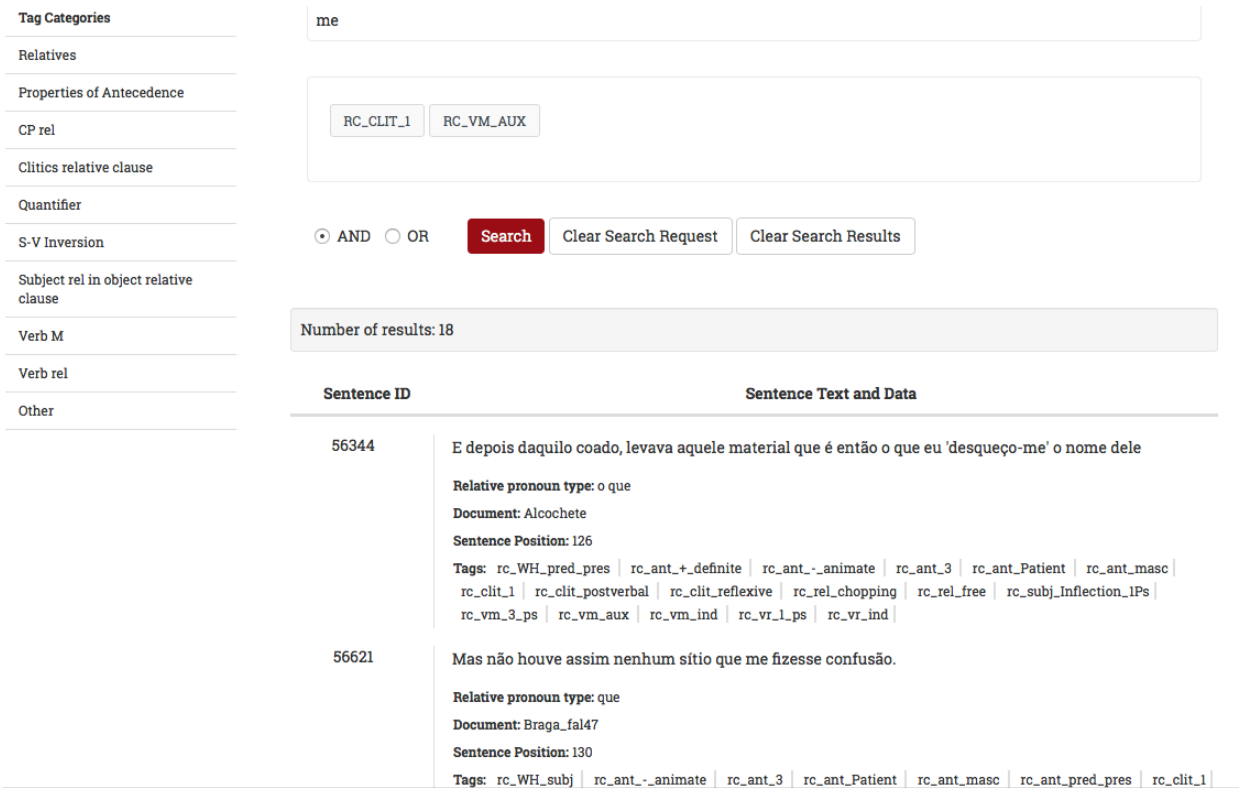


Figure 3: Screenshot of the search result page of the Web Application designed for the DFG-Projekt PO 1642/1-1.

A Digital Library of Grammatical Resources for European Dialects. In Maristella Agosti, et al., editors, *Digital Libraries and Archives*, volume 249 of *Communications in Computer and Information Science*, pages 61–74. Springer.

Agosti, M., Alber, B., Nunzio, G. M. D., Dussin, M., Rabanus, S., and Tomaselli, A. (2012). A Curated Database for Linguistic Research: The Test Case of Cimbrian Varieties. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Benincà, P. and Poletto, C. (2007). The ASIS Enterprise: A View on the Construction of a Syntactic Atlas for the Northern Italian Dialects. *Nordlyd. Monographic issue*

*on Scandinavian Dialects Syntax*, 34(1).

Bird, S., Klein, E., and Loper, E. (2009). *Natural Language Processing with Python*. O'Reilly Media, 1st edition, July.

Chiarcos, C. (2012). Interoperability of Corpora and Annotations. In Christian Chiarcos, et al., editors, *Linked Data in Linguistics - Representing and Connecting Language Data and Language Metadata*, pages 161–179. Springer.

Di Buccio, E., Di Nunzio, G. M., and Silvello, G. (2014). A linked open data approach for geolinguistics applications. *IJMSO*, 9(1):29–41.

Goebl, H. (2010). Dialectometry: theoretical prerequisites, practical problems, and concrete applications (mainly with examples drawn from the "Atlas linguistique de France" 1902-1910). *Dialectologia*, 1:63 – 77.

König, E. and Mengel, A. (2000). Linguistic Databases, John Nerbonne, Ed. *Journal of Logic, Language and Information*, 9(4):513–517, October.

Peter Buneman and James Cheney and Wang-Chiew Tan and Stijn Vansummeren. (2008). Curated Databases. In *Proceedings of the Twenty-seventh ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, PODS '08, pages 1–12, New York, NY, USA. ACM.

Soria, C., Calzolari, N., Monachini, M., Quochi, V., Bel, N., Choukri, K., Mariani, J., Odijk, J., and Piperidis, S. (2014). The language resource strategic agenda: the flarenet synthesis of community recommendations. *Language Resources and Evaluation*, 48(4):753–775.

Spärck Jones, K. (2007). Computational Linguistics: What About the Linguistics? *Computational Linguistics*, 33(3):437–441, September.

Wieling, M. and Nerbonne, J. (2015). Advances in dialectometry. *Annual Review of Linguistics*, 1(1):243–264.